# White Paper

Report ID: 107433

Application Number: HT-50070-12

Project Director: Trevor Munoz

Institution: University of Maryland, College Park

Reporting Period: 10/1/2012-9/30/2015

Report Due: 12/31/2015

Date Submitted: 8/31/2016

PROJECT WHITE PAPER

# Digital Humanities Data Curation

HT-50070-12

**Institutes for Advanced Topics in the Digital Humanities**
**National Endowment for the Humanities**


***Project director:***

**Trevor Muñoz**
Assistant Dean for Digital Humanities Research,
University Libraries,
Associate Director,
Maryland Institute for Technology in the Humanities,
University of Maryland

**Local project directors:**

**Julia Flanders**
Head, Digital Scholarship Group,
Professor of the Practice of English,
Northeastern University

**Megan Senseney**
Project Coordinator for Research Services
Center for Informatics Research in Science and Scholarship,
Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign

The Digital Humanities Data Curation (DHDC) Institute was awarded an Institutes for Advanced Topics in the Digital Humanities grant by the National Endowment for the Humanities in the amount of $248,721. DHDC was a collaborative initiative led by the Maryland Institute for Technology in the Humanities (MITH) in cooperation with the Women Writers Project at Northeastern University and the Center for Informatics Research in Science and Scholarship at the University of Illinois Graduate School of Library and Information Science (GSLIS). The institute was designed to serve as an opportunity for humanities scholars with all levels of expertise—from beginners to the most advanced—to receive guidance in understanding the role of data curation in enriching humanities research projects. Institute workshops intended not only to further the educational efforts of the selected participants but also to allow for the adaptation of data curation curricula from the several degree and advanced certificate programs in research data curation within existing library and information science graduate programs to the specific needs of the digital humanities research community. Ultimately the goal of DHDC was to create a community of practitioners invested in humanities data curation from a range of different disciplinary communities including digital humanities, information science, and digital libraries.

# Project Activities

#### ACTIVITIES FROM OCTOBER 1, 2012 TO SEPTEMBER 30, 2013

*Technical development on the DH Curation Guide.* The DH Curation Guide (http://guide.dhcuration.org) is a community-based curricular resource, providing an annotated, contextualized listing of key resources for understanding data curation in the humanities. The Guide was first developed at the University of Illinois as part of a project dedicated to extending data curation research and curriculum with library and information graduate programs to the humanities. This earlier project was funded by the Institute for Museum and Library Services (IMLS). For DHDC, the Guide served as a course reader base and was also considered as a mode of wider dissemination for the knowledge that is shared within the institute events (see discussion below). During this performance period, the project team shifted the Guide's hosting from Illinois to Maryland and also transitioned to a simpler platform that will reduce the overall time expended for programming and content management. A web page announcing the first DHDC institute was also added to the existing DH Curation site (see http://dhcuration.org/institute).

*Curriculum Development and Workshop Planning*. Due to Carole Palmer's planned sabbatical in Spring 2013, the project team began curricular development earlier than anticipated in Fall 2012, in order to benefit from Palmer's data curation expertise and prior experience in developing and running the IMLS-funded Summer Institutes in Data Curation from 2008–2011. Trevor Muñoz, Julia Flanders, and Dorothea Salo took the lead in planning a three-day curriculum for DHDC. Each day of workshop comprised lecture-style presentations by session leaders, case studies, hands-on activities, and group discussions. For the first

institute, Dr. Ted Underwood, Associate Professor of English at the University of Illinois Urbana-Champaign, agreed to share a local case study presenting his experiences and personal challenges addressing data curation issues related to his research, which includes running a variety of algorithmic approaches against large-scale text corpora for literary history analysis.

All key planning activities for the first workshop occurred as planned and on schedule, including making local arrangements for meeting space and housing, creating an institute page on the DH Curation website, issuing a call for applications, reviewing applications, and notifying participants of their acceptance. The project team had not, however, anticipated the overwhelming positive response to our call for applications. In total, we received 111 applications for our first institute, which is limited to 20 participants. In addition to institute applications, another 225 people signed up to receive more information about future workshops. The first round of applicants' demographics ranged from local to international with 18 faculty members, 45 library and information science professionals, 7 alternative academics, 36 graduate students, and 5 others. With an 18% acceptance rate, the DHDC team was able to be highly selective in the choice of participants.

*Workshop #1*. The first DHDC workshop was held at Graduate School of Library and Information Science at the University of Illinois, from June 24–26, 2013. Palmer opened the first workshop with remarks about humanities data curation in the context of the data curation education program at GSLIS. The first day of the institute consisted of providing an introduction to humanities data curation, an interactive exercise in which participants introduced their data as well as themselves, a discussion of the social dimensions of data, and an activity introducing participants to different approaches to data management planning. Day two began with a discussion of the nature of data and digital objects followed by a tour of data curation systems and platforms. In the afternoon, Underwood presented the case of his own research in the digital humanities for a discussion of real-world curation issues and an exercise in problem solving. Day two closed with a discussion of the role collections play in curation activities. The third and final day of the workshop opened with a curation activity using data from the New York Public Library's "What's on the Menu" project and Google Refine. The institute concluded with a set of sessions devoted to legal and policy aspects of data curation, risk assessment and mitigation, and sustainability.

The three-workshop format of DHDC allowed the project team to better understand the needs of the community, assess how workshops might better meet those needs, and revise curriculum accordingly. To that effect, the project team rigorously documented the first institute and analyzed materials generated from workshop activities, note-taking, social media interactions, and direct feedback via an evaluation survey (see below). The project team began making in-development resources available through the DHDC Workshop Wiki on GitHub, including copies of all presentation slides and workshop notes

(https://github.com/digital-humanities-data- curation/dhdc-workshop/wiki). In an effort both to foster ongoing conversation among workshop participants and to create potential avenues of participation for applicants who were not selected, the team experimented with a closed pilot of an online discussion forum (see additional discussion below).

*Curriculum Revisions*. In response to participants' requests for more hands-on activities, the project team made several revisions to the DHDC curriculum after the first workshop. Revisions included an introduction to depositing and retrieving items in an Islandora repository; an exercise in exploring the many functional roles and intersections of professionals who participate in data curation activities; and a brainstorming activity in which participants shared examples of personal solutions to data curation problems that they've already implemented in their own work or at their home institution. Additional curriculum revisions included developing a lecture on metadata; tweaking minor aspects of pre-existing sessions; further developing the scaffolding for previously-introduced exercises; and building in more flexible time for technical troubleshooting and professional networking.

*Planning for Workshop #2*. Once again, the project team was overwhelmed by the enthusiastic response to the call for applications to the second DHDC institute at the University of Maryland. In total, DHDC received 136 applications for the second institute, which could only accommodate 20 participants. The total number of subscribers to the Digital Humanities Data Curation email list by summer 2013 was over 500 people. The selected cohort for the second institute included 2 faculty, 8 library and information science professionals, 4 alternative academics, 4 graduate students, and 2 others from a range of institutions across the United States.

#### ACTIVITIES FROM OCTOBER 1, 2013 TO SEPTEMBER 30, 2014

*Workshop #2*. The second workshop was held from October 16–18, 2013, at the University of Maryland, College Park. In addition to curriculum revisions discussed above, the instructors included additional time for advanced topic sessions based on attendees' interests. The workshop participants identified two topics: "Open Data and Data Journalism" and "Linked Data", which were covered on the last day of the workshop. Kari Kraus, Associate Professor in the College of Information Studies and the Department of English at the University of Maryland provided a case study from her recent research on the second day of the workshop.

*Curriculum Revisions, Round Two*. In response to participants' enthusiasm for the attendee-driven session on linked data at the second workshop, the topic was formalized and incorporated into the DHDC curriculum. The data personae and self-profiling exercises were replaced with a birds-of-a-feather breakout session based on topics that arose during participant introductions and a mind mapping exercise to 1) determine what data curation encompasses and 2) explore the data curation landscape. The instructors also explored a new approach to the session on data management planning, which included peer review of

real data management plans and a breakout session in which participants brainstormed ways to think beyond project-based plans to customize and improve the process of planning for data management across varying scales and scenarios.

*Workshop #3.*

Once again, the project team was overwhelmed by the enthusiastic response to the call for applications to the third and final DHDC workshop at Northeastern University. In total, the project received 112 applications for the second institute, which could only accommodate 21 participants. This round of applicants' demographics ranged from local to international with 21 faculty members, 47 library and information science professionals, 8 alternative academics, 25 graduate students, and 11 others. With a 19% acceptance rate, DHDC was able to be highly selective in the choice of participants. The final selected cohort included 5 faculty, 8 library and information science professionals, 4 alternative academics, 3 graduate students, and 1 other from a wide range of institutions. The third workshop was held from April 30–May 2, 2014, at Northeastern University.

The other significant change to conduct of the third workshop (aside from those described above) involved the handling of the digital humanities case study. For this institute's case study, the instructors recruited representatives from two very different DH projects, both located at Northeastern University. The case study presenters were given more thorough context for the role of the session within the broader workshop, and the two-project structure facilitated a comparative approach to participant discussions in considering how different projects may pose discrete curatorial challenges and require different approaches. Jim McGrath, Ph.D. candidate in English at Northeastern University, presented Our Marathon: The Boston Bombing Digital Archive and WBUR Oral History Project, and Elizabeth Dillon, professor of English and co-director of the NULab for Texts, Maps, and Networks at Northeastern University, presented The Early Caribbean Digital Archive.

### ACTIVITIES FROM OCTOBER 1, 2014 TO SEPTEMBER 30, 2015

*Additional dissemination activities*. In fall 2014, the National Endowment for the Humanities approved a one-year no-cost extension for the DHDC project through September 30, 2015. The project used the extension period to continue refining curriculum, explore additional opportunities for dissemination of project resulting, and develop the project's final performance reports.

The third cohort of workshop participants were the first group invited to test the pilot of the discussion forum initiated after Workshop #2. Participants used the space for introductions prior to the workshop as well as for giving and receiving advice about data-related problems and for sharing tools, techniques, and resources. Ultimately however, uptake of the discussion forums was very limited. Many digital humanities practitioners

already participate in numerous online fora, from Twitter to Digital Humanities Question and Answers, to email discussion lists. The project team determined that an additional forum devoted to humanities data curation was unlikely to be sustainable and the pilot project was closed down in Spring 2015.

After an internal review and discussions with potential hosts, the project team decided likewise not to expand the content of the DH Curation Guide or invest further resources in its development. The Guide will remain online as an open access resource. The introductory article on "Humanities Data Curation" by Flanders and Muñoz has been cited 10 times since 2013, in published literature in both the humanities and information science. From anecdotal reports, the DHDC team is aware of several courses in information science graduate programs using materials from the Guide in courses. Moreover, given the limited number of resources specifically devoted to data curation for the humanities, the Guide site often ranks highly in Google search results. However, to expand the Guide to include new content would require substantial investment of time for soliciting submissions, managing editorial workflows, and supporting digital publication. The resources to support this investment do not currently exist at any of the potential host institutions the project team considered. The Guide, which predates the DHDC project, served as a teaching resource and was considered as an outlet for dissemination of project results. Given the decision not to invest further in the Guide as an active publication, the usefulness to DHDC for dissemination is greatly decreased.

One of the most successful outcomes of this project has been sustainably embedding DHDC curricular content within recurring, self-sustaining digital humanities training initiatives. In summer 2014, Senseney organized a workshop entitled "Data Curation and Access for the Digital Humanities" at the Digital Humanities Oxford Summer School (DHOxSS) with colleagues from the HathiTrust Research Center, the Center for Informatics Research in Science and Scholarship, the Bodleian Library, and the Oxford eResearch Center. Her sessions drew upon DHDC lectures, activities, and exercises related to data management, using Open Refine, and discussion-based case studies with the goal of extending the impact of the DHDC curriculum to the additional audiences. Senseney led a revised version of the workshop entitled "Humanities Data: Curation, Analysis Access and Reuse" at the 2015 Digital Humanities Oxford Summer School with a colleague the Oxford eResearch Center. As part of the Humanities Intensive Learning + Teaching (HILT) Institute in 2015, Muñoz collaborated with Katie Rawson from the University of Pennsylvania to offer an advanced course, "Humanities Data Curation Praxis," which experimented with more tool-intensive curricular materials and with guided opportunities for participants to workshop projectspecific data curation strategies in a small group setting.

At several points during the no-cost extension period, team members shared insights from DHDC through invited talks and guest appearances in other training workshops. For

example, Muñoz led mini workshops on humanities data curation for the Center for Digital Humanities at Princeton University (April 2015), as part of the Early Modern Digital Agendas Institute (EMDA) held by the Folger Institute (June 2015), and as part of the "bootcamp" for postdoctoral fellows (July 2015) supported by the Council on Library and Information Resources (CLIR). These additional opportunities for dissemination introduced additional audiences to data curation principles, tools, and methods.

# Accomplishments

The DHDC project succeeded in introducing over 60 humanities scholars, librarians and other information professionals, graduate students, and technologists to the basic concepts of data curation as well as to several useful tools and methods for maintaining the value of humanities research over time through three in-person workshops hosted at the University of Illinois, Urbana-Champaign, the University of Maryland, and Northeastern University. Furthermore, the large response to the calls for participation in these workshops—far beyond what could be accommodated—identified a clear need for additional training and activities focused on data curation for the humanities.

# Audiences

One of the stated goals of the DHDC project was to address a distinct shortage of focused training opportunities for working professionals that address the discipline-specific data curation training needs of digital humanities scholars and the librarians or other information specialists who collaborate closely with digital humanists. The selected participants reflected this goal. Library and information professionals represented the largest group of participants, likely reflecting their higher awareness of data curation issues. Graduate students and young scholars represented the next largest cohort, followed by faculty members, whose time and participation is often harder to secure. The DHDC institute also served a number of unaffiliated researchers, representatives of federal government agencies and professional organizations, and one member of local government.

# Evaluation

The DHDC project conducted an evaluation survey after each of the three institute workshops and used the results of these surveys to adjust and improve the curriculum.

Eleven out of 20 participants (55%) responded to the post-institute evaluation survey for the first workshop. All respondents considered themselves as having beginner-level or moderate experience with data curation. Participants generally agreed or strongly agreed that they gained a greater understanding of data curation in context of digital humanities research (90.9%) and a greater understanding of how to make data curation decisions related to

creating, organizing, using, and preserving digital content (72.7%). However, only 45.5% of respondents agreed or strongly agreed that indicated that they had gained a better understanding of tools for comparison and analysis of humanities data, indicating an area for improvement when reviewing the curriculum for the second institute. Responses to open-ended questions expressed enthusiasm for the institute's instructors and overall framework ("Loved Trevor's framework and overall themes for humanities data."; "I love the presenters! They did an awesome job for a new subject that has very few hard and fast rules") while also recommending the addition of more hands-on activities ("I would focus less on data curation theory and incorporate more "real life" examples into the workshop."; "I wish we could use 1-2 tools with our own data to see how they work with some guided practice.")

Eleven out of 20 participants (55%) responded to the post-institute evaluation survey for the second workshop. Of those who responded, 72% considered themselves as having beginner-level or moderate experience with data curation. This represented an increase in experienced or expert-level participants from the first to the second institute, a decision that the workshop organizers made while selecting workshop participants with the hope that such participants would help seed and enliven workshop discussions. When asked to evaluate individual sessions, participants identified the newly-developed metadata lecture as most valuable with 90% (n = 10) of respondents ranking the lecture as "Very valuable", the highest rank on a 5-point Likert scale. Other well-received sessions included an introductory lecture on conceptual frameworks for humanities data curation, a lecture on understanding the nature of digital objects, and the newly-developed exercise on "sharing what works" with 80% (n = 10) of respondents ranking these sessions as "Very valuable". Participants were least enthusiastic about a data personae exercise, a self-profiling data exercises, and the case study presentation. Upon reflection, the workshop organizers agreed that the case study required more advanced consultation with the presenter and session scaffolding for the participants. The other two-exercises were new to the DHDC workshops and were not used again.

Overall, participants were pleased with the general balance of lectures, hands-on work with tools and technologies, and small-group exercises with a general preference for slightly more emphasis on hands-on work. In open-ended responses to a question seeking additional comments about the workshop curriculum, two respondents highlighted a preference for a more in-depth and comprehensive session on linked data. Topics recommended for future DHDC workshops included:

- More time spent on useful curation tools and more time discussing the common data types that are outcomes of DH projects;
- Add-ins and plug-ins that enhance the curatorial function of commonly used applications, mind map tools, and workflow systems and frameworks;
- Metadata schemas for specific fields and disciplines;

- Case studies more tightly aligned with representative DH projects; and
- A dedicated session on linked open data.

All 10 respondents indicated that they were either very satisfied (70%) or satisfied (30%) with the training received at the DHDC workshop, and 80% of respondents confirmed that they would definitely recommend DHDC to a colleague.

Eleven out of 21 participants (52%) responded to the post-institute evaluation survey for the third workshop—consistent with the rate from the previous two events. Of those who responded, 27% self identified as experienced or expert in data curation compared with 73% with beginner level or moderate experience. This demographic spread is consistent with the second institute and informed by experiences from the first institute, after which the workshop organizers chose to include a larger selection of more experienced data curators with the goal of seeding and enlivening workshop discussions across the board. When asked to evaluate individual sessions, participants identified the lecture introducing data management plans at the most valuable with 100% (n=9) of respondents ranking the lecture as “Very valuable”, the highest rank on a 5-point Likert scale. Other well-received sessions included a lecture on understanding the nature of digital objects, an exercise based on identifying the significant properties of the NYPL “What’s on the Menu?” dataset, and a lecture on metadata and linked data with 78% (n=9) of respondents ranking these sessions as “Very valuable”. Participants were least enthusiastic about the hands-on Islandora repository exercise, the two case studies, and a group exercise on customizing and improving plans for data management. Upon reflection, these responses (combined with lessons learned from previous institutes) underscore the challenges associated with case-based approaches to data curation instruction, the additional scaffolding required for successfully negotiating hands-on sessions with complex tools intended to be used as part of routine institutional (rather than project-based) practice. Also tensions between data management as a core component of technical projects and the data management plan as a conceptual exercise and prerequisite for funding remained a challenging balance to get right in a workshop setting.

Overall, participants were pleased with the general balance of lectures, hands-on work with tools and technologies, and small-group exercises with a general preference for slightly more emphasis on hands-on work with tools and technologies over small-group exercises and discussions. In open-ended responses to a question seeking additional comments about the workshop curriculum, one respondent requested, “more explanation of how the different layers of information architecture work together.” Other comments touched upon the community building aspects of the program with varying degrees of enthusiasm ranging from “Great training and teaching methods and formats for engagement, group ownership, learning and community building” to “it was difficult to find a balance in groups between it being broad enough to involve everyone but specific enough to be useful.”

There were fewer recommendations for future topics from the third workshop cohort than from previous cohorts, however, two themes emerged: data management plans and collected best practices. Two respondents suggested an exercise in which participants draft new data management plans. Notably, an analogous session was an early component of the institute curriculum, but the organizers evolved in a different direction based on feedback from prior institutes. This divergent response from the third workshop might be a consequence of including more experienced practitioners (who felt ready to tackle writing data management plans). Another possible explanation could be increased awareness of the data management plan requirements from NEH and other funders. Three respondents requested more emphasis on best practices, suggesting possible future directions as we prepare online resources for the digital humanities community.

Nearly all respondents indicated that they were either very satisfied (56%) or satisfied (33%) or with the training received at the DHDC workshop, and 67% of respondents confirmed that they would definitely recommend DHDC to a colleague. Final comments from respondents included:

- “I had a great time and I feel I learned a lot. That metaphor for ‘seeing’ things differently and being able to apply what we have learned to our future research is a good lesson for all. I feel I will approach my datasets in a more nuanced and intelligent way in the future. Thank you! Great organization and material.”
- “I really appreciated the broad mix of professions and disciplines represented. That was of enormous value to me in understanding the broader concepts and how they relate to one another.”

Taken together these evaluations suggest a number of conclusions about the state of data curation knowledge in the humanities and about the future needs of the field. One intervention the project team expected to make was to respond directly to the new data management plan requirements for projects funded by NEH and other funders. In most of the workshops, participants were grappling with data curation theory as well as with the large number of participants, institutions, and other actors involved in managing research information over time. Thus, it seemed as though participants struggled to feel sufficient mastery of these topics to effectively distill their curation strategies into the artefact (a two-page plan) required by funding applications. Iterations of the data management planning sessions which reduced the engagement with the specific funder guidelines in favor of a wider exploration of the expectations and possibilities for data management planning seemed more successful based on participant evaluations. Also, participants felt more confident workshopping existing data management plans rather than generating new plans—suggesting the value of model and sample plans. Topics such as linked data were introduced in response to participant feedback. Other topics such as risk assessment were introduced earlier in the workshop schedule to provide context but remained relatively unchanged, while yet other topics such as

sustainability were increasingly truncated (probably due to the need for more extensive treatment than time allowed). Guided discussion and small group work grew over the series of workshop while individual work and lectures were reduced. Ultimately, the evaluations trace the project team's experimentation with different potential elements of humanities data curation training—a process which would probably continue in any future extension to the DHDC project.

# Continuation of the Project

The project team discussed two possible continuations of the DHDC project. First, as the number of people working in the area of humanities data curation grows, it might be valuable to convene a summit meeting—or even another full institute—focused on curriculum and pedagogy for humanities data curation. The explicit aim of this activity would be to develop and harmonize training opportunities that cross the boundaries of formal graduate training, either in the humanities or information science. In other words, what might a coherent approach to teaching data curation look like as both modules or units within larger curricula and as extracurricular or in-service training outside formal degree programs? Second, because of the humanities particular engagement with the study of underrepresented communities, a data curation institute series which incorporated standards for managing traditional cultural knowledge (such as those practices developed by the Mukurtu project) or which incorporated a particular focus on disciplines such as Black Studies, Latinx Studies, or LGBTQ studies might be a valuable continuation of the DHDC work.

DHDC project team members expect to remain active in promoting and teaching humanities data curation, however, the team has no current plans to pursue additional activities.

# Long Term Impact

The DHDC project will have an impact beyond the period of performance in two ways—through the cohort of workshop participants and through the development of a pool of experienced teachers of humanities data curation. Informally, participants have reported using the concepts and tools learned in DHDC workshops in their digital humanities scholarship and teaching. These reports often mention two specific components. First, participants report relying on the definition of data developed through the workshops. This definition of data—as information serving in the role of evidence for knowledge claims—they report, resonates well with other humanities scholars, even those who might be sceptical of digital tools and methods. Second, Open Refine, as a "power tool" for data inspection and reconciliation, remains popular with many workshop attendees.

Through the incorporation of humanities data curation courses into ongoing training activities such as the Digital Humanities Oxford Summer School and HILT, DHDC team members have partnered with other instructors to continue offering training opportunities based on the institute curriculum. In this way, the pool of instructors and leaders for humanities data curation activities continues to grow.

# Grant Products

The project maintains a website at http://www.dhcuration.org, which includes both the DH Curation Guide and documentation of DHDC workshops through schedules and slide decks.

The project team presented on preliminary findings from DHDC at the Digital Humanities 2014 conference:

> Senseney, M., Muñoz, T., Flanders, J., & Fenlon, A. (2014). Digital Humanities Data Curation Institutes: Challenges and preliminary findings. Poster presented at Digital Humanities, Lausanne, Switzerland, July 812, 2014.